An analysis of pulsation periods of long-period variable stars

Jeffrey D. Hart

Department of Statistics, Texas A&M University
Collaborators

Chris Koen – South African Astronomical Observatory, Cape, South Africa

Fred Lombard – Rand Afrikaans University, Auckland Park, South Africa
Outline

• Brief introduction to variable stars
• Problem of interest
• Statistical model
• Investigating assumptions
• Characteristics of fitted models
• Testing for trend
• Conclusions
Variable stars

- Characterized by brightness changes over time

- **Long period variables** – distinguished by substantial brightness changes

- Changes are roughly sinusoidal with typical periods between 100 and 300 days.

- The period of a given star is determined by its internal structure.

- *Period changes important as they reflect changing physical conditions in the stars.*
Data

- Database of 378 long-period variables
- Times of maximum brightness recorded by amateur and professional astronomers.
- Data processed by American Association of Variable Star Observers (AAVSO).

**AAVSO website:** [http://www.aavso.org](http://www.aavso.org)

- For each star in the database, we consider $P_1, \ldots, P_n$, the observed times between successive maximum brightnesses.
- $n$ ranges from 32 to 212 with median 74.
Light curve and periods
Typical data plots

Omicron Ceti

R Aquilae

R Bootis

R Camelo
Problem of interest

- For each star in the database, determine if there is evidence of a trend in the times between consecutive maximum brightnesses.

- Formally, we wish to test the null hypothesis

\[ H_0 : E(P_1) = E(P_2) = \cdots = E(P_n). \]

Will conduct a test of this hypothesis for each of the 378 stars in the database.

Variety of methods and challenges

1. Time series and spectral analysis

2. Data-driven model selection

3. Smoothing

4. Wild failure of asymptotic distribution theory for a likelihood ratio statistic

5. Bootstrap in an unconventional way to deal with 4 and computational issues.
Model

- $P_1, \ldots, P_n$: observed lengths of time between successive maximum brightnesses for a given star

- Model used by astronomers:

  $$P_j = P + T_j + I_j + \epsilon_j - \epsilon_{j-1}, \quad j = 1, \ldots, n$$

  - $P$: long-run average of all $P_j$s
  - $T_1, \ldots, T_n$: constants representing trend
  - $I_1, \ldots, I_n$: “intrinsic” errors, i.e., random deviations intrinsic to the star
  - $\epsilon_1, \ldots, \epsilon_n$: errors made in determining times of maximum brightness
Hypothesis of interest

In terms of our statistical model, the hypothesis of interest is

\[ H_0 : T_1 = T_2 = \cdots = T_n = 0. \]

We wish to test this hypothesis for each star.
Error series

Usual assumptions in astronomy literature:

- $I_1, \ldots, I_n$ are i.i.d. $(0, \sigma_I^2)$.
- $\epsilon_0, \ldots, \epsilon_n$ are i.i.d. $(0, \sigma_\epsilon^2)$.
- The two series are independent.
Error series

Let $\xi_j = I_j + \epsilon_j - \epsilon_{j-1}$, $j = 1, \ldots, n$.

- $\{\xi_j\}$ has same covariance function as first order moving average process.
- First lag correlation of $\{\xi_j\}$ is always between $-1/2$ and 0.
- Spectral density of $\{\xi_j\}$ is monotone increasing, or “high frequency.”
De-trending of data

Each data set is de-trended as follows:

- Fit a fifth degree polynomial to \((\frac{j - 1/2}{n}, P_j)\),
  \(j = 1, \ldots, n\).

- Compute residuals

\[ e_j = P_j - \hat{P}_j, \quad j = 1, \ldots, n. \]

These residuals will be analyzed in various ways to discover properties of the error process.
Estimated log-spectra for five stars

These are the stars whose estimated spectra at 0 frequency were at the 10th, 25th, 50th, 75th, and 90th percentiles among all 378 stars.
Investigating assumptions

To perform a valid test of $H_0$, the error series

$$I_j + \epsilon_j - \epsilon_{j-1}, \quad j = 1, \ldots, n,$$

needs to be modeled correctly.

Questions to consider:

- Is the error series homoscedastic?
- Is it reasonable to assume that the two series $\{I_j\}$ and $\{\epsilon_j\}$ are Gaussian?
Homoscedasticity

Absolute residuals for six variable stars and local linear smooths
Model for heteroscedasticity

- We assume heteroscedasticity arises only from increasing precision in determining times of maximum brightness.

- We thus assume intrinsic errors are i.i.d.

- Will assume measurement errors have the form

\[ \epsilon_j = \exp \left[ \left( \frac{\beta_0}{2} \right) + \left( \frac{\beta_1}{2} \right) \left( \frac{j - 1/2}{n} \right) \right] \eta_j, \quad j = 1, \ldots, n, \]

where \( \beta_0 \) and \( \beta_1 \) are unknown constants and \( \eta_1, \ldots, \eta_n \) are independent and identically distributed with mean 0 and variance 1.
Normality

- Is it reasonable to assume that the errors are normally distributed?

- Standardized residuals:

\[ E_i = \frac{e_i}{\hat{\sigma}_i}, \quad i = 1, \ldots, n, \]

where \( e_1, \ldots, e_n \) are the residuals from a fitted fifth degree polynomial and \( \hat{\sigma}_i^2, \quad i = 1, \ldots, n \), is a local linear smooth of \( e_1^2, \ldots, e_n^2 \).

- The standardized residuals for all 378 stars were pooled together to investigate whether there are departures from normality.
Distribution of pooled residuals

Let $\epsilon_{ij}$ be i.i.d. $f_i$ for $j = 1, \ldots, m$, $i = 1, \ldots, n$.

Kernel estimate of pooled data:

$$\hat{f}_h(x) = \frac{1}{nmh} \sum_{i=1}^{n} \sum_{j=1}^{m} K \left( \frac{x - \epsilon_{ij}}{h} \right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \hat{f}_h(x; i).$$
Distribution of pooled residuals

Suppose that for a fraction $w$ of the $n$ data sets, $f_i \equiv \phi$, the standard normal density. Then $\hat{f}_h(x)$ estimates

$$w\phi(x) + (1 - w)g_n(x),$$

where $g_n$ is a mixture of nonnormal densities.

Nonnormality should be detectable from $\hat{f}_h$ if

- The fraction $1 - w$ of nonnormal densities is fairly substantial, and
- the density $g_n$ is substantially nonnormal.

Presumably, the latter would happen if the departures from normality tended to be similar to each other.
Kernel density estimate of pooled residuals

An analysis of pulsation periods of long-period variable stars – p.22
Final model

1. Trend modeled as a polynomial of degree $m$, which is unknown but not larger than 15.

2. The two series $\{I_j\}$ and $\{\epsilon_j\}$ are independent of each other.

3. Intrinsic errors $\{I_j\}$ follow an AR(1) process such that $E(I_j) = 0$ and $\text{Var}(I_j) = \sigma_I^2$, $j = 1, \ldots, n$.

4. The experimental errors are independent Gaussian random variables with variance structure as described earlier.

5. All parameters (for given $m$) estimated by maximum likelihood.
Estimating polynomial degree

- Polynomial degree is an important model parameter since no trend hypothesis is equivalent to $m = 0$.

- Pokta (2004) shows that BIC tends to select too large an $m$ when $\sigma_I^2 = 0$.

- We use a modified BIC that uses a nonuniform prior on $m$:

  $$P(m = 0) = 1/2 \quad \text{and} \quad P(m = k) \propto k^{-2}, \quad k = 1, \ldots, 15.$$
Distribution of estimated polynomial degrees

Modified BIC chose degree 0 for 291, or 77%, of the 378 stars.

Conditional distribution of $\hat{m}$ given $\hat{m} > 0$
Distribution of estimates of error parameters

- “correlation” – AR(1) parameter for intrinsic series
- “standard deviation” – standard deviation of $P_j$ at most current observation time
- “ratio” – ratio of experimental error standard deviations at latest and earliest observation times
- “R” – ratio of intrinsic variance to measurement error variance at earliest observation time
Distribution of estimates of error parameters

![Graphs showing distributions of correlation, standard deviation, ratio, and R parameters.](image-url)
Parameters of measurement error variance

\[ \text{Var}(\epsilon_j) = \exp(\beta_0 + \beta_1 x_j) \]
Standard deviation of data vs. $R$

$$R = \frac{\sigma_I^2}{\exp(\beta_0)}$$
Testing for trends

- Perform likelihood ratio test of no-trend hypothesis.

- Use statistic

\[ T = 2 \log(\hat{L}_6 / \hat{L}_0), \]

where \( \hat{L}_j \) is the maximized likelihood for a \( j \)th degree polynomial model.

- For computational reasons, this statistic preferred to omnibus lack-of-fit statistics as in Aerts, Claeskens and Hart (1999).
Bootstrap

- $\chi^2_6$ approximation to null distribution of $T$ is extremely poor.
- The null is composite. Determining distribution of $T$ as function of unknown parameters and $n, \ldots$, nightmarish.
- Bootstrap the Bayesian/frequentist testing procedure of Bayarri and Berger (2000), \ldots, priceless.
Bayarri-Berger $p$-value

1. $\theta$: vector of parameter values left unspecified under the null hypothesis.

2. $\pi$: prior distribution for $\theta$ over parameter space $\Theta$

3. $f(p; \theta)$: joint probability distribution of $P_1, \ldots, P_n$ under $H_0$ for a given value $\theta$ of the parameter vector
Bayarri-Berger $p$-value

The $p$-value for observed value $t_{\text{obs}}$ of $T_6$ is

$$p = \text{Prob}(T_6 \geq t_{\text{obs}}),$$

where the probability is defined with respect to the marginal distribution $g$ of $P_1, \ldots, P_n$, i.e.,

$$g(p) = \int_{\Theta} f(p; \theta) \pi(\theta) \, d\theta.$$

We take $\pi$ to be the actual distribution of parameter values over the population of stars.

We then bootstrap by sampling randomly and with replacement from the set of 378 stars.
Bootstrap algorithm

1. A star is randomly selected from the set of 378. Let $\hat{R}$ and $\hat{\beta}_1$ denote the estimates of $R$ and $\beta_1$ for the chosen star at its BIC optimal polynomial degree.

2. Generate sample

$$P_j^* = \hat{R}^{1/2} I_j^* + \exp(\hat{\beta}_1 x_j / 2) \epsilon_j^* - \exp(\hat{\beta}_1 x_{j-1} / 2) \epsilon_{j-1}^*, \quad j = 1, \ldots, n,$$

where $I_1^*, \ldots, I_n^*, \epsilon_0^*, \ldots, \epsilon_n^*$ are i.i.d. $N(0, 1)$ and $x_j = (j - 1/2)/n, \ j = 0, 1, \ldots, n$.

3. Compute $T_{6}^*$ for data generated in 2.

4. Repeat steps 1-3 independently 1000 times for a given $n$. 
Distribution of $T_6$ as a function of $n$

Let $n = 68$ and $T = 12.59$.

large sample $p$-value = 0.05  actual $p$-value $\approx 0.46$
Distribution of $P$-values

An analysis of pulsation periods of long-period variable stars – p.36
Testing multiple hypotheses

- Should take into account performing multiplicity of tests.

**Q** is proportion of rejected hypotheses that are falsely rejected. FDR is $E(Q)$.

- Let $p_1 \leq p_2 \leq \cdots \leq p_N$ be ordered $p$-values from $N$ independent tests. Define $k$ to be the largest $i$ such that

  $$p(i) \leq \frac{i}{N} \alpha.$$  

- If $H_i$ is the null hypothesis corresponding to $p(i)$, $i = 1, \ldots, N$, then procedure that rejects $H_i$ iff $i \leq k$ ensures that the FDR is no more than $\alpha$. 
35 and 19 stars have significant trends when controlling FDR at 5% and 1% levels, respectively.
$P$-value as a function of mean period
\[ \log(SNR) \] as a function of mean period
Astronomy conclusions

- Convincing evidence that the observed pulsation periods of most long period variables are heteroscedastic.

- Using a method that controls the false discovery rate to be 0.05, 35 stars have significant trends in times between maximum brightness.

- Most of the trends that are statistically significant have little upward or downward tilt, but rather a wavelike behavior.

- There is a clear tendency for strength of trend to be positively related to the mean period of a star.
Statistics points

- Testing for nonnormality in many data sets simultaneously using pooled residuals.

- Theoretical investigation of bootstrapping to approximate Bayarri-Berger $p$-values.

- Develop strategy for optimizing FDR and TDR (true discovery rate) – analog of size and power in the “testing many hypotheses” problem.